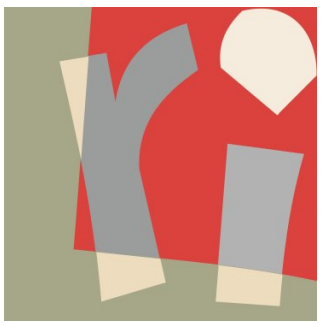
Uncertain Data Management
Sources of Uncertain Data

Antoine Amarilli[1], Silviu Maniu[2]

[1] Télécom ParisTech

[2] LRI

November 21st, 2016



Uncertain Data Management

Database systems usually assume that data is correct and complete

Uncertain Data Management

Database systems usually assume that data is correct and complete

- Incomplete and missing data
- Imprecise data
- Noisy data
- Untrustworthy data

Uncertain Data Management

Database systems usually assume that data is correct and complete

- Incomplete and missing data
- Imprecise data
- Noisy data
- Untrustworthy data

→ Which applications produce uncertain data nowadays?

Never-Ending Language Learning

NELL: Read the Web

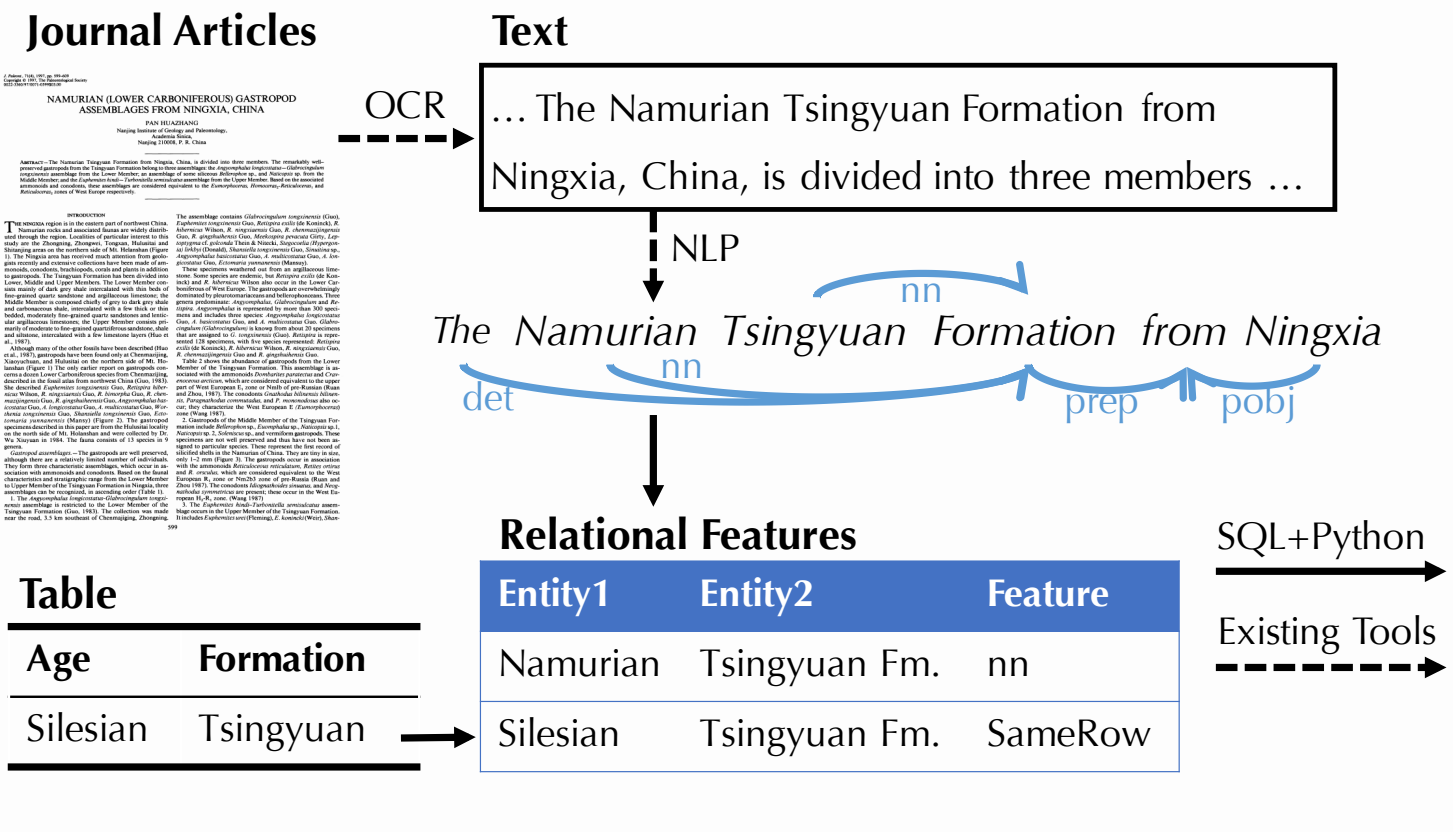
Recently-Learned Facts

Refresh

instance	iteration	date learned	confidence
kampioenschap van zwitserland is a sports race	955	20-oct-2015	95.0
cochran mill nature center is an aquarium	955	20-oct-2015	96.9
kozy shack chocolate pudding is a kind of candy	956	23-oct-2015	90.3
red delicious apple tree is a plant	955	20-oct-2015	92.8
sale miami dade county is a sport	955	20-oct-2015	99.1
chicken001 eat black beans	955	20-oct-2015	100.0
wrigley field is the home venue for the sports team chicago cubs	959	07-nov-2015	100.0
lorena ochoa is a person who has residence in the geopolitical location mexico	958	03-nov-2015	100.0
umass lowell river hawks hired john calipari	955	20-oct-2015	98.4
nuggets participated in the event games	955	20-oct-2015	100.0

Information extraction

DeepDive: extract facts from journal articles



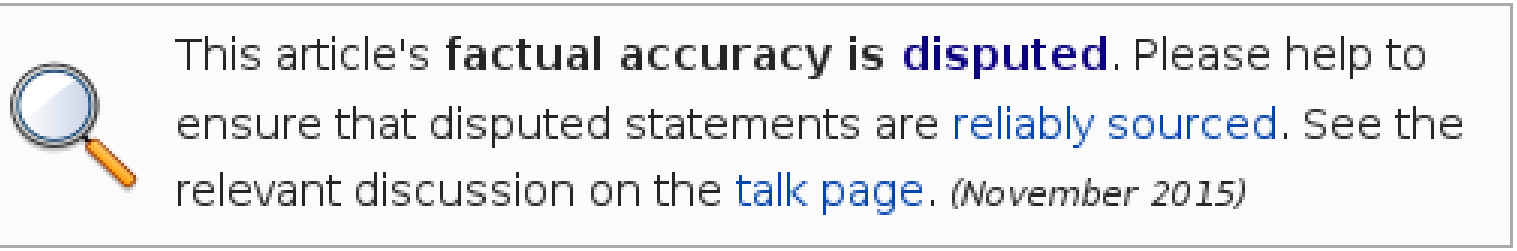
Many sources of uncertainty

- Errors in sources:

This article's **factual accuracy is disputed**. Please help to ensure that disputed statements are reliably sourced. See the relevant discussion on the talk page. *(November 2015)*

Many sources of uncertainty

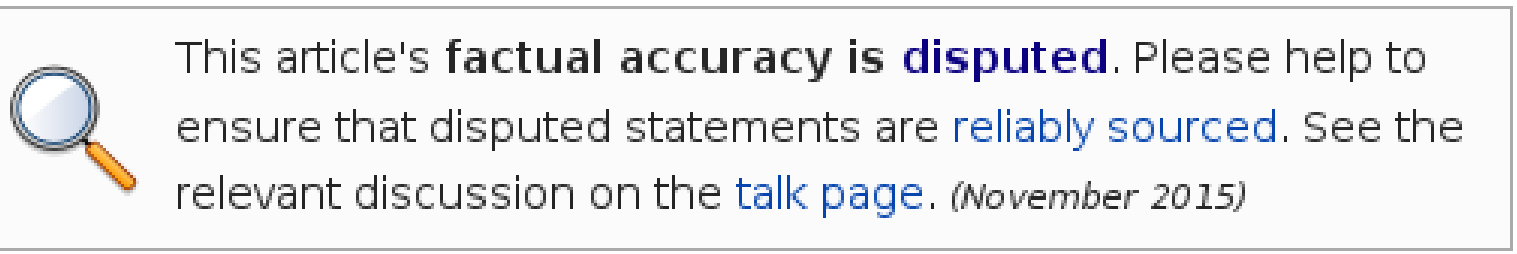
- Errors in sources:



- Entity disambiguation:
  "The place and function of Venus in Ovid..."
  "Computed backscattering function of Venus and the moon..."

Many sources of uncertainty

- Errors in sources:



- Entity disambiguation:
  "The place and function of Venus in Ovid..."
  "Computed backscattering function of Venus and the moon..."
- Natural language parsing:
  "such cities as New York", "such cities as Mohenjo-Daro"
  "the classification of such cities as urban"

Many sources of uncertainty

- Errors in sources:

This article's **factual accuracy is disputed**. Please help to ensure that disputed statements are reliably sourced. See the relevant discussion on the talk page. *(November 2015)*

- Entity disambiguation:
  "The place and function of Venus in Ovid..."
  "Computed backscattering function of Venus and the moon..."
- Natural language parsing:
  "such cities as New York", "such cities as Mohenjo-Daro"
  "the classification of such cities as urban"
- Anaphora resolution:
  "Obama told Hollande that he was not a spying target"

Many sources of uncertainty

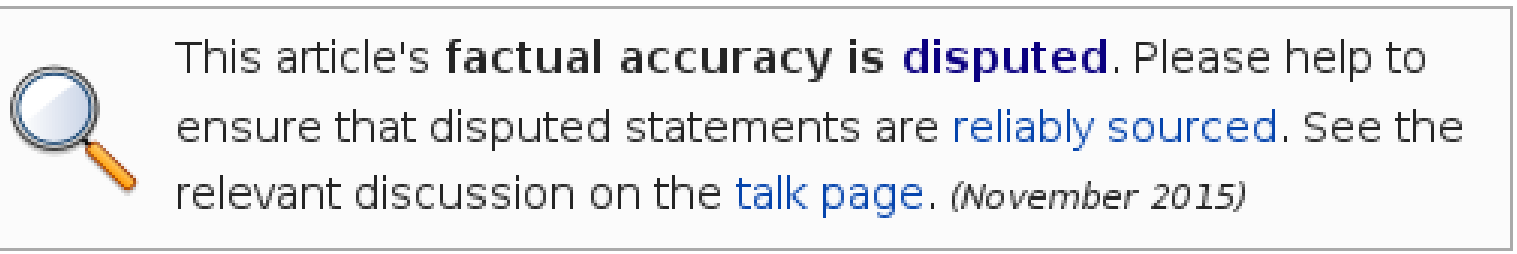
- Errors in sources:

This article's **factual accuracy is disputed**. Please help to ensure that disputed statements are reliably sourced. See the relevant discussion on the talk page. *(November 2015)*

- Entity disambiguation:
  "The place and function of Venus in Ovid..."
  "Computed backscattering function of Venus and the moon..."
- Natural language parsing:
  "such cities as New York", "such cities as Mohenjo-Daro"
  "the classification of such cities as urban"
- Anaphora resolution:
  "Obama told Hollande that he was not a spying target"
- Noise:
  - performance using reinforcement learning is a machine-learning algorithm

Many sources of uncertainty

- Errors in sources:



- Entity disambiguation:
  "The place and function of Venus in Ovid..."
  "Computed backscattering function of Venus and the moon..."
- Natural language parsing:
  "such cities as New York", "such cities as Mohenjo-Daro"
  "the classification of such cities as urban"
- Anaphora resolution:
  "Obama told Hollande that he was not a spying target"
- Noise:
  - performance using reinforcement learning is a machine-learning algorithm 
- Incompleteness

Crowdsourcing

Amazon Mechanical Turk

All HITs

1-10 of 2751 Results

Sort by: HITs Available (most first) GO! Show all details | Hide all details 1 2 3 4 5 › Next » Last

Transcribe data — View a HIT in this group

Requester: p9r **HIT Expiration Date:** Nov 18, 2015 (23 hours 59 minutes) **Reward:** $0.03

Time Allotted: 45 minutes

Description: Please transcribe the data from the following images

Keywords: transcribe, handwriting, data entry

Qualifications Required:

HIT approval rate (%) is greater than 90

Classify Receipt — View a HIT in this group

Requester: Jon Brelig **HIT Expiration Date:** Nov 24, 2015 (6 days 23 hours) **Reward:** $0.02

Time Allotted: 20 minutes

Description: Looking at a receipt image, identify the business of the receipt

Keywords: image, receipt, categorize, transcribe, extract, data, entry, transcription, text, easy, qualification, jon, brelig, prod

Qualifications Required:

Total approved HITs is not less than 1000

HIT approval rate (%) is not less than 97

Location is US

Crowdsourcing

Amazon Mechanical Turk

All HITs

1-10 of 2751 Results

Sort by: HITs Available (most first) GO! Show all details | Hide all details 1 2 3 4 5 › Next » Last

Transcribe data — View a HIT in this group

Requester: p9r **HIT Expiration Date:** Nov 18, 2015 (23 hours 59 minutes) **Reward:** $0.03

Time Allotted: 45 minutes

Description: Please transcribe the data from the following images

Keywords: transcribe, handwriting, data entry

Qualifications Required:

HIT approval rate (%) is greater than 90

Classify Receipt — View a HIT in this group

Requester: Jon Brelig **HIT Expiration Date:** Nov 24, 2015 (6 days 23 hours) **Reward:** $0.02

Time Allotted: 20 minutes

Description: Looking at a receipt image, identify the business of the receipt

Keywords: image, receipt, categorize, transcribe, extract, data, entry, transcription, text, easy, qualification, jon, brelig, prod

Qualifications Required:

Total approved HITs is not less than 1000

HIT approval rate (%) is not less than 97

Location is US

→ Users are untrustworthy!

Sentiment analysis

n	Most positive n-grams	Most negative n-grams
1	engaging; best; powerful; love; beautiful	bad; dull; boring; fails; worst; stupid; painfully
2	excellent performances; A masterpiece; masterful film; wonderful movie; marvelous performances	worst movie; very bad; shapeless mess; worst thing; instantly forgettable; complete failure
3	an amazing performance; wonderful all-ages triumph; a wonderful movie; most visually stunning	for worst movie; A lousy movie; a complete failure; most painfully marginal; very bad sign
5	nicely acted and beautifully shot; gorgeous imagery, effective performances; the best of the year; a terrific American sports movie; refreshingly honest and ultimately touching	silliest and most incoherent movie; completely crass and forgettable movie; just another bad movie. A cumbersome and cliche-ridden movie; a humorless, disjointed mess
8	one of the best films of the year; A love for films shines through each frame; created a masterful piece of artistry right here; A masterful film from a master filmmaker,	A trashy, exploitative, thoroughly unpleasant experience ; this sloppy drama is an empty vessel.; quickly drags on becoming boring and predictable.; be the worst special-effects creation of the year

→ Possible mistakes!

Schema mappings

	Possible Mapping	Prob
$m_1 =$	{(pname, name), (email-addr, email), (current-addr, mailing-addr), (permanent-addr, home-addr)}	0.5
$m_2 =$	{(pname, name), (email-addr, email), (permanent-addr, mailing-addr), (current-addr, home-addr)}	0.4
$m_3 =$	{(pname, name), (email-addr, mailing-addr), (current-addr, home-addr)}	0.1

(a)

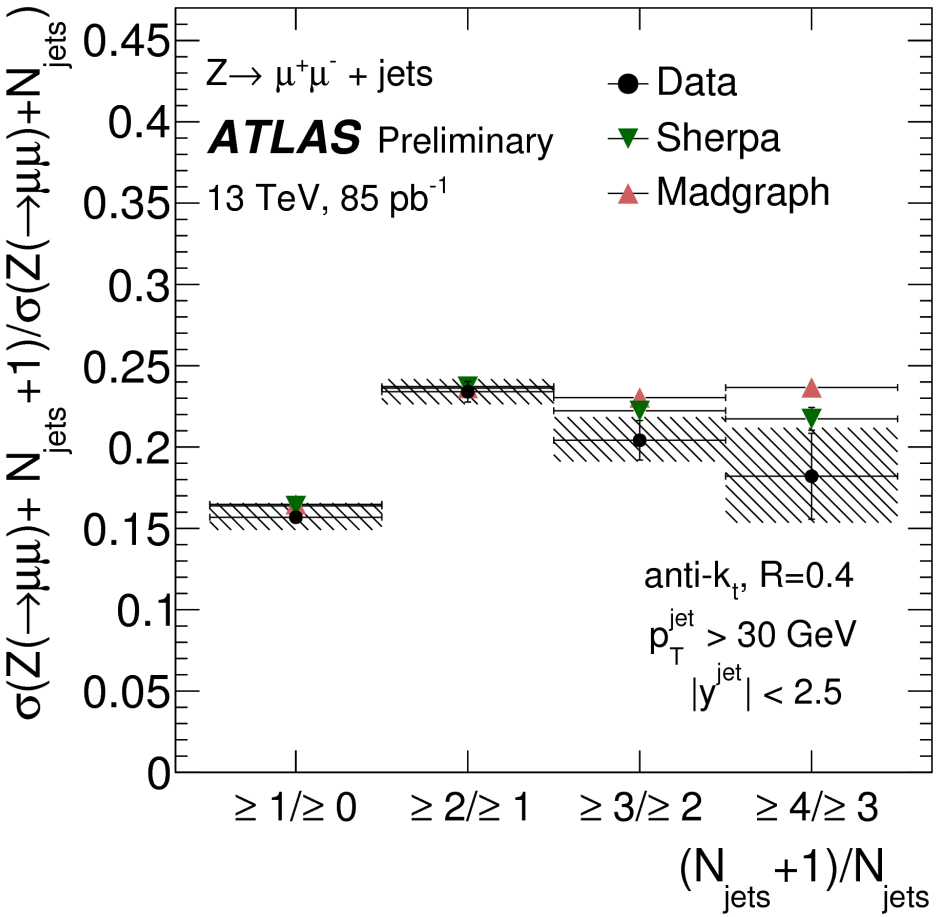
*p*name	*e*mail-addr	*c*urrent-addr	*p*ermanent-addr
Alice	alice@	Mountain View	Sunnyvale
Bob	bob@	Sunnyvale	Sunnyvale

(b)

Tuple (mailing-addr)	Prob
('Sunnyvale')	0.9
('Mountain View')	0.5
('alice@')	0.1
('bob@')	0.1

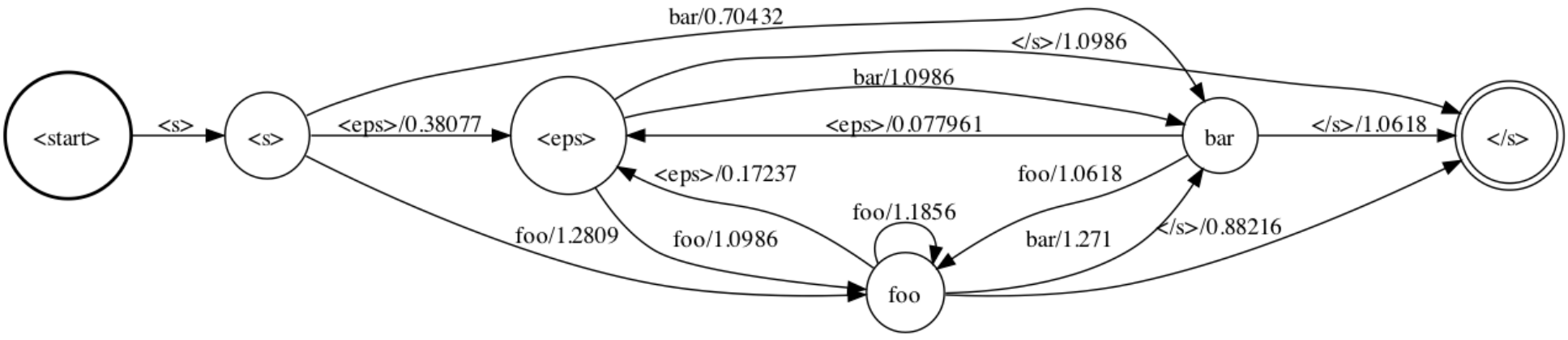
(c)

Scientific data



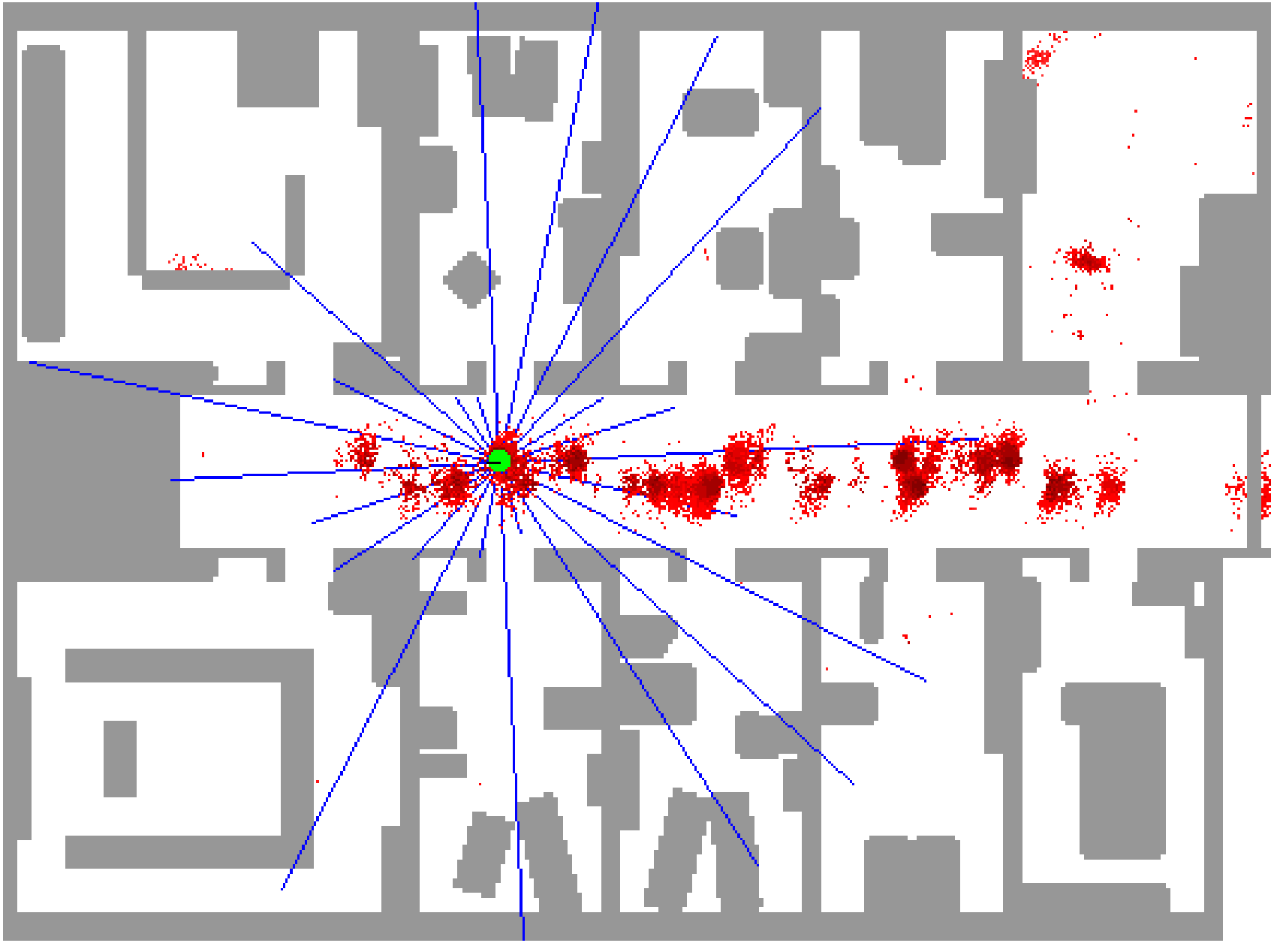
→ Measurement errors

Speech recognition and OCR



→ Decoding output is uncertain

Robotics



Other applications

- Data integration: combine data across sources
- Data cleaning: fix errors in stale/outdated data
- Machine learning: predictions are uncertain
- Data mining: trends extracted from large datasets
- Computational biology: genomic data management

... and much more!

Image Credits

- Slide 5: http://rtw.ml.cmu.edu/
- Slide 7: https://en.wikipedia.org/wiki/Template:Disputed
- Slide 6: [Zhang, 2015], page 9
- Slide 13: https://www.mturk.com/
- Slide 15: [Socher et al., 2013], page 10
- Slide 16: [Dong et al., 2009], page 4
- Slide 17: https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/CONFNOTES/ATLAS-CONF-2015-041/fig_06b.png
- Slide 18: https://code.google.com/p/transducersaurus/wiki/CascadeTutorial
- Slide 19: https://www.cs.washington.edu/robotics/mcl/

References I


Dong, X. L., Halevy, A., and Yu, C. (2009).
Data integration with uncertainty.
The VLDB Journal—The International Journal on Very Large Data Bases, 18(2):469–500.


Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C., Ng, A., and Potts, C. (2013).
Recursive deep models for semantic compositionality over a sentiment treebank.
In *Proc. EMNLP*.

References II


Zhang, C. (2015).
DeepDive: A Data Management System for Automatic Knowledge Base Construction.
PhD thesis, University of Winconsin–Madison.
https://cs.stanford.edu/people/czhang/zhang.thesis.pdf.